

# Proximity penalty priors for Bayesian mixture models

Matthew Sperrin

*Department of Mathematics and Statistics, Lancaster University, UK*

*m.sperrin@lancaster.ac.uk*

July 28, 2011

## Abstract

When using mixture models it may be the case that the modeller has a-priori beliefs or desires about what the components of the mixture should represent. For example, if a mixture of normal densities is to be fitted to some data, it may be desirable for components to focus on capturing differences in location rather than scale. We introduce a framework called proximity penalty priors (PPPs) that allows this preference to be made explicit in the prior information. The approach is scale-free and imposes minimal restrictions on the posterior; in particular no arbitrary thresholds need to be set. We show the theoretical validity of the approach, and demonstrate the effects of using PPPs on posterior distributions with simulated and real data.

Keywords: Bayesian; Identifiability; MCMC; Mixture Model; Prior Specification.

## 1 Introduction

Mixture models are widely recognized as a useful tool for inference in a variety of settings. Having been first used over 100 years ago (for example, in Pearson, 1894), more recently mixture models are enjoying a revival, thanks to advances in computational methods for inference. In particular, the EM algorithm (Dempster et al., 1977) and MCMC (see, for example, Diebolt and

Robert, 1994) have driven considerable advances in the field. See McLachlan and Peel (2000) for a general overview of mixture models; Fruhwirth-Schnatter (2006) provides an overview of Bayesian mixture models, which are the focus of this paper.

We recall the definition of a mixture model and introduce notation. Suppose  $n$  observations,  $y_1, \dots, y_n$ , are taken from a  $K$ -component mixture distribution where all the components have the same distributional form, with mixture-specific parameters  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K)$ , global parameters  $\boldsymbol{\eta}$  and mixing weights  $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$ , summarised by  $\boldsymbol{\gamma} = (\boldsymbol{\pi}, \boldsymbol{\theta}, \boldsymbol{\eta})$ . The mixture distribution for a single observation  $Y_i$  is then given by

$$g(y_i|\boldsymbol{\gamma}) = \sum_{k=1}^K \pi_k f_k(y_i|\boldsymbol{\theta}_k, \boldsymbol{\eta}), \quad (1)$$

with  $K \geq 1$ ,  $\pi_k > 0$  ( $k = 1, 2, \dots, K$ ),  $\sum_{k=1}^K \pi_k = 1$  and  $f_k(\cdot|\boldsymbol{\theta}_k, \boldsymbol{\eta})$  is a density function parametrised by  $\boldsymbol{\theta}_k$  and  $\boldsymbol{\eta}$ .

A Bayesian approach to estimating the parameters of the mixture distribution of Equation (1) involves the specification of priors for the parameters  $\boldsymbol{\gamma}$ . The issue of prior specification in this context has a number of difficulties.

First, fully improper priors cannot be used for component-specific parameters in mixture models, since doing so causes the posterior to be improper also (see, for example, McLachlan and Peel, 2000). However, proper priors, even with large variance, can have considerable influence on the posterior distribution, and the extent of this influence can be difficult to assess (Marin et al., 2005). Re-parametrisation in a hierarchical manner and allowing only the global parameters to be improper is one solution: this is considered by Mengersen and Robert (1996), and Roeder and Wasserman (1997). Another possibility is to use data-dependent priors, as considered by Richardson and Green (1997), and Wasserman (2000).

Second, where no component specific information is available, identical priors may be proposed for the components of each parameter. This leads to a non-identifiable posterior, which is known as the label switching problem. This has been well studied (see, for example Stephens, 2000; Jasra et al., 2005; Sperrin et al., 2010, and references therein).

Third, constructing independent priors for component parameters may not be sensible, as the components only have meaning relative to one another (Lee et al., 2008).

This third issue is the focus of this paper. We consider in detail the idea that priors should be specified relative to each other. We introduce a strategy for doing so that we call ‘proximity penalty priors’ (PPPs). The basic idea is that priors are specified in two parts: first, each prior is specified independently, corresponding to standard existing approaches; second, a proximity penalty is applied, which penalises the joint prior distribution of certain configurations of parameters. We show that the construction makes theoretical sense.

Section 2 introduces the idea of PPPs. Section 3 illustrates the consequences of the PPP approach on real and simulated data; the paper concludes with a discussion in Section 4.

## 2 Proximity Penalty Priors

We begin with a simple result that establishes the validity of the PPP approach.

**Proposition 1.** *Suppose the prior for  $\gamma$ , given by  $p(\gamma)$ , can be separated as*

$$p(\gamma) = p_1(\gamma)p_2(\gamma).$$

*Denote the likelihood by  $L(\gamma)$  and the posterior by  $q(\gamma)$ , so that  $q(\gamma) \propto L(\gamma)p(\gamma)$ . Suppose that a new parameter vector  $\gamma^*$  can be simulated from a proposal distribution  $r(\gamma^*) = L(\gamma^*)p_1(\gamma^*)$ , and the existing value of  $\gamma$  is  $\gamma^m$ . Then if we set*

$$\gamma^{m+1} = \begin{cases} \gamma^* & \text{with probability } \min\left(1, \frac{p_2(\gamma^*)}{p_2(\gamma^m)}\right) \\ \gamma^m & \text{otherwise,} \end{cases} \quad (2)$$

*the result is equivalent to a Metropolis-Hastings update.*

*Proof.* The acceptance probability for the Metropolis-Hastings procedure with proposal density  $r(\cdot)$  and posterior  $q(\cdot)$  is

$$\min\left(1, \frac{q(\gamma^*)r(\gamma^m)}{q(\gamma^m)r(\gamma^*)}\right).$$

Substituting in these densities gives the result. □

In the context of this work the portion of the prior  $p_1(\cdot)$  corresponds to the independent specification of the parameters, for which standard distributions could be used; the portion  $p_2(\cdot)$  corresponds to the novel part of the prior that jointly assesses the values of the parameters and penalises undesirable combinations.

Suppose that the priors  $p_1(\cdot)$  are conjugate. Then an MCMC approach would proceed, on each iteration, by generating proposed new parameters according to a Gibbs sampling scheme with the full conditionals based on the prior component  $p_1(\cdot)$ , then accepting the proposed parameters according to a Metropolis Hastings ratio on the prior component  $p_2(\cdot)$ .

We illustrate the idea with an example. Consider a mixture of two normal distributions

$$p(y_i|\boldsymbol{\gamma}) = \pi_1 N(y_i; \mu_1, \sigma_1^2) + \pi_2 N(y_i; \mu_2, \sigma_2^2), \quad (3)$$

with  $\pi_1 + \pi_2 = 1$ , and all the parameters  $\boldsymbol{\gamma} = (\pi_1, \pi_2, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$  unknown. Standard conjugate prior choices would then be a Dirichlet distribution for the pair  $(\pi_1, \pi_2)$ , normal distributions for  $\mu_1$  and  $\mu_2$ , and inverse-gamma distributions for  $\sigma_1^2$  and  $\sigma_2^2$ . Throughout this paper we will use the empirical Bayes prior distributions suggested by Richardson and Green (1997) unless otherwise stated. We may believe a-priori that the key difference between the two components is the location. If the components are not well separated or the amount of data is small it is important that such prior information is captured. By Proposition 1, we can reflect these beliefs in a separate part of the prior  $p_2(\cdot)$ . A sensible such choice is

$$p_2(\boldsymbol{\gamma}) = |\mu_1 - \mu_2|. \quad (4)$$

Such a function assigns more prior weight to larger differences between  $\mu_1$  and  $\mu_2$ . In isolation, the above  $p_2(\cdot)$  is improper but provided  $p_1(\cdot)$  is proper the overall prior is proper. Such a prior enjoys scale invariance in the sense that  $p_2(a\mathbf{x}_1)/p_2(a\mathbf{x}_2) = p_2(\mathbf{x}_1)/p_2(\mathbf{x}_2)$  for all non-zero  $a$ . This may or may not be desirable. An alternative would be to specify a distance  $\delta$  as a minimum distance between  $\mu_1$  and  $\mu_2$ , i.e.

$$p_2(\boldsymbol{\gamma}) = \mathbf{1}_{(|\mu_1 - \mu_2| > \delta)}.$$

This generates the question of how  $\delta$  should be specified, but may be appropriate in some situations.

More generally, for a mixture distribution with  $K$  parameters, suppose there exists a component-specific parameter  $\phi_k$  for each component  $k = 1, \dots, K$ , and the difference between the components is a-priori believed (or, from the point of view of model interpretation, desired) to be in terms of this parameter. Then we propose setting

$$p_2(\gamma) = \min_{k \neq l} |\phi_k - \phi_l|. \quad (5)$$

On the other hand, for a mixture distribution with  $K$  parameters, if there exists a component-specific parameter  $\psi_k$  for each component  $k = 1, \dots, K$ , and each component is a-priori expected or desired to have *similar* values of this parameter, we could set

$$p_2(\gamma) = \max_{k \neq l} |\psi_k - \psi_l|^{-1}. \quad (6)$$

Here, the scale free nature of  $p_2(\cdot)$  is an advantage in that we do not have to quantify ‘similar’. More generally,  $p_2(\gamma)$  could be constructed as any multiplicative combination of Equations (5) and (6). The procedure can also be applied when the number of components  $K$  is allowed to vary, in which case it makes sense only within fixed values of  $K$  in the same way that the label switching problem only has meaning within fixed values of  $K$  (Nobile and Fearnside, 2007).

## 3 Examples

### 3.1 Mixture of Two Normals

Our first illustration takes the simple mixture of two normals example. We generate 100 observations from the density given in Equation (3), with  $\mu_1 = 0$ ,  $\mu_2 = 2$ ,  $\sigma_1^2 = \sigma_2^2 = 1$  and  $\pi_1 = \pi_2 = 0.5$ . We consider two prior specifications:

- (a) the standard specification given in Richardson and Green (1997), denoted *without PPP*;
- (b) a two part prior  $p(\gamma) = p_1(\gamma)p_2(\gamma)$ , with  $p_1(\gamma)$  as given in Richardson and Green (1997) and  $p_2(\gamma)$  as given in Equation (4), denoted *with PPP*.

In both cases we fix the number of components  $K = 2$ . In (b), we are therefore adding an explicit prior opinion that the difference between the two components is in the locations  $\mu_1$  and  $\mu_2$ .

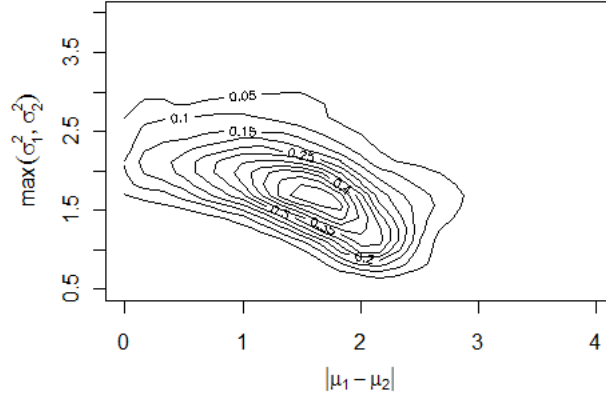
Figure 1 compares a bivariate projection of the posterior onto the absolute difference  $|\mu_1 - \mu_2|$  and  $\max(\sigma_1^2, \sigma_2^2)$  without and with the PPP. Without the PPP, posterior mass is assigned to the situation where  $|\mu_1 - \mu_2|$  is small and  $\max(\sigma_1^2, \sigma_2^2)$  is large. This corresponds to a case where a mixture distribution with similar means but different variances is fitted. In Figure 2 we see that such a mixture is well supported by the data (dashed line in the figure). Once the PPP is applied, far less posterior mass is assigned to this scenario, since our prior distribution specifically tells us to exclude such cases.

Figure 3 gives the marginal bivariate posterior of  $(\mu_1, \mu_2)$ , with and without the PPP. Without the PPP, the posterior appears to have a single mode at approximately  $\mu_1 = \mu_2 = 1$ ; with the PPP, the posterior is bimodal with modes at approximately  $(\mu_1 = 0, \mu_2 = 2)$  and  $(\mu_1 = 2, \mu_2 = 0)$ . The bimodality in the PPP case is a consequence of label switching; if component-specific inference is required, post-hoc relabelling should be carried out (see, for example, Sperrin et al., 2010). The unimodality in the non PPP case is caused by the two means being very close together and the variances to differ, corresponding to a different interpretation of the mixture components.

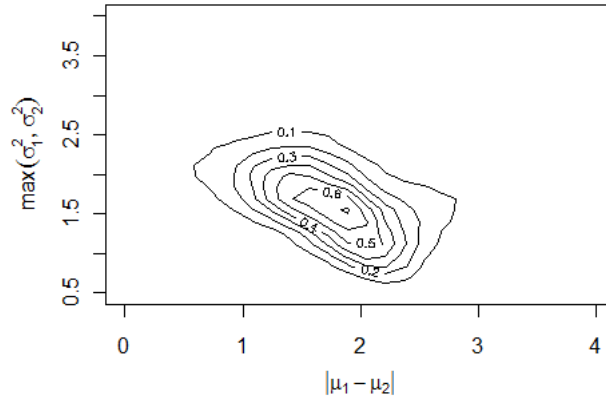
We also ran the same comparison without assuming a fixed number of components  $K$  (using the birth-death method of Stephens, 2000), putting a  $\text{Poisson}(1)$  prior distribution on the number of components  $K$  (see Nobile and Fearnside, 2007, for a justification of the use of this prior). Similar results to the above were observed when we looked at the output conditional on  $K = 2$ .

## 3.2 Galaxy Data

The galaxy dataset is commonly used to illustrate mixture modelling techniques (see Jasra et al., 2005, for a recent investigation of this dataset in the mixture modelling context). Briefly, it consists of the velocities of 82 galaxies, but the velocities appear to cluster, suggesting different groups of galaxies that we may wish to identify (see Figure 4). If we model these data using a mixture, it is likely that we wish our mixture components to represent the clusters with different mean velocities, hence the PPP of Equation (5) could be considered in this scenario. We run a variable dimension sampler with the details as above, with normally distributed components assumed and a  $\text{Poisson}(1)$  prior distribution on the number of components  $K$ . We



(a) without PPP



(b) with PPP

Figure 1: Posterior contour plots of  $|\mu_1 - \mu_2|$  versus  $\max(\sigma_1^2, \sigma_2^2)$

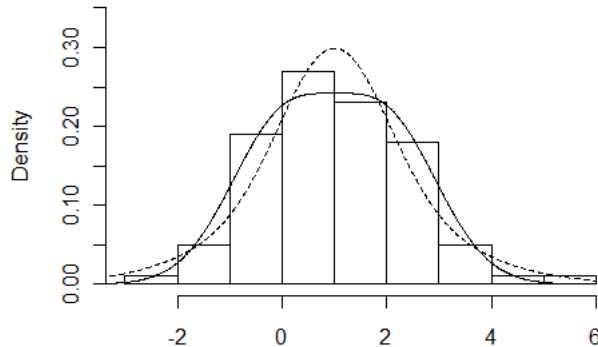


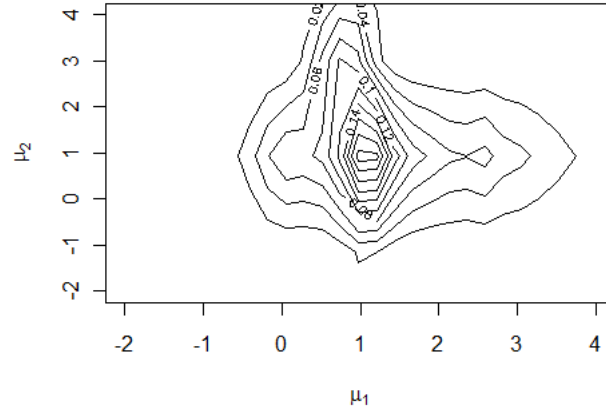
Figure 2: Histogram of 100 realisations from  $0.5N(0, 1) + 0.5N(2, 1)$  with true density overlaid (solid line) and alternative density,  $0.5N(1, 1) + 0.5N(1, 4)$  also overlaid (dashed line)

compare the results of standard priors (i.e. those given in Richardson and Green, 1997) with the standard priors plus the PPP. Both with and without the PPP, the values of  $K$  with the majority of posterior support are  $K = 3$  and  $K = 4$  (but see Aitkin, 2001, for discussion on the posterior of the number of components in a mixture model). For the  $K = 3$  case the posterior means are already well separated, and the PPP has little or no effect on the posterior means. We look in more detail at the  $K = 4$  case.

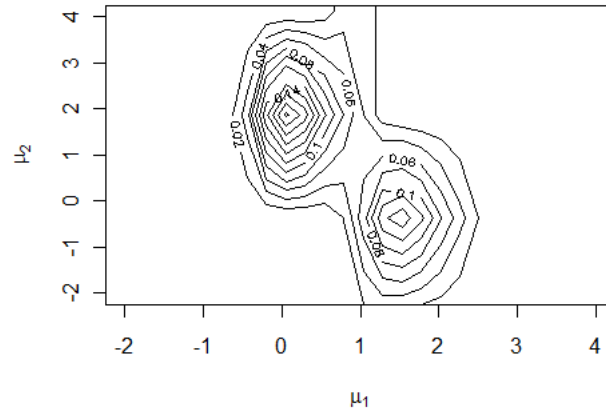
In order to avoid the label switching issue, we first consider the posterior of a generic  $\mu_k$  without relabelling, estimating this by combining into a single vector all samples from the posterior  $\mu_k$ , for  $k = 1, 2, 3, 4$ , conditional on  $K = 4$ . We can do this since invariance of the posterior under re-parametrisation means we can ignore the labels. The resulting density plot is given in Figure 5. The interesting difference to note here is that with the PPP four distinct peaks can be observed in the density, whereas without the PPP the middle two peaks cannot be distinguished. This does, however, depend on the smoothing parameter used in the non-parametric density estimate.

To consider this further we mitigate the label switching issue by applying





(a) without PPP



(b) with PPP

Figure 3: Posterior contour plots of  $\mu_1$  versus  $\mu_2$

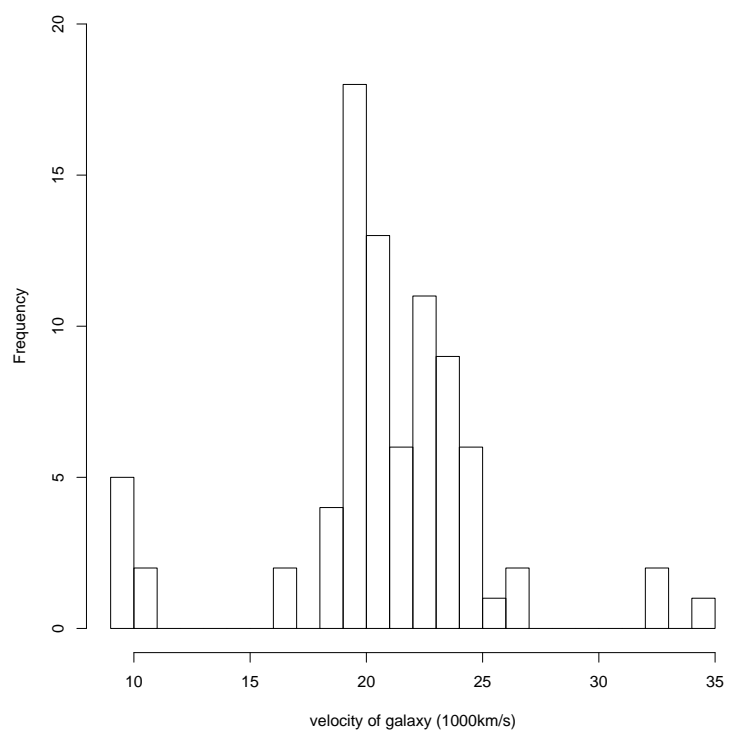


Figure 4: Histogram of the velocities of 82 galaxies

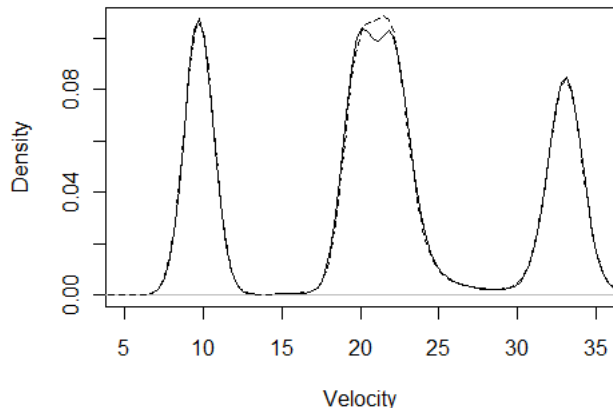


Figure 5: Smoothed density of a generic  $\mu_k$  for the galaxy data. Without PPP: dashed line; with PPP: solid line.

the identifiability constraint  $\mu_1 < \mu_2 < \mu_3 < \mu_4$ , then look at the posterior density of  $(\mu_3 - \mu_2)$ . This is given in Figure 6. We see that applying the PPP causes more separation between the two component means (less mass at small differences).

## 4 Discussion

In this paper we have introduced the idea of incorporating weak joint information about parameters in a mixture model into the prior specification. In particular we have introduced proximity penalty priors (PPPs) as a method of explicitly declaring an a-priori opinion (or interest) in components that differ on a certain parameter. The formulation is designed to allow this opinion to be as vague as possible: we avoid making any statement about the magnitude of the difference that should be observed between the components, i.e. the method is scale-free.

With the focus of this paper being introduction of the idea, the examples were kept fairly simple. The idea, however, is very general and could be applied in more complex models. For example, in an application such as genetics we may wish to construct a mixture of regressions with many co-

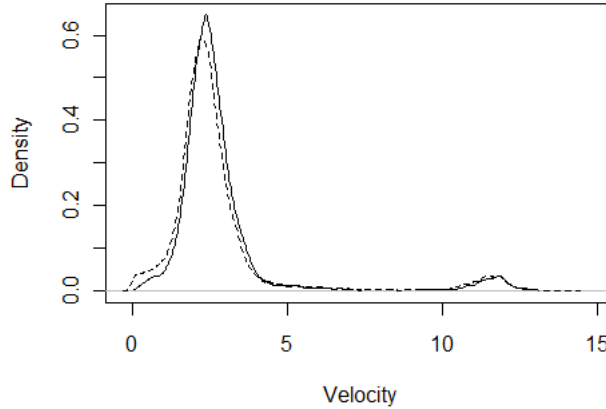


Figure 6: Smoothed density of  $(\mu_3 - \mu_2)$  for the galaxy data with  $K = 4$  after an IC is applied. Without PPP: dashed line; with PPP: solid line.

variates. Suppose there are  $p$  covariates and  $K$  mixtures, with the coefficient of the  $j^{\text{th}}$  covariate in the  $k^{\text{th}}$  mixture given by  $\beta_{jk}$ . Then we could consider the PPP

$$p_2(\gamma) = \max_j \min_{k \neq l} |\beta_{jk} - \beta_{jl}|,$$

to reflect a belief that each component should have at least one coefficient that differs from the value in every other component.

Another potential extension is to replace the  $L_1$ -norm assumed in the PPP with an  $L_s$ -norm, i.e. considering a generalisation of, for example, Equation (4), to

$$p_2(\gamma) = |\mu_1 - \mu_2|^s.$$

In this generalised setting, we note that  $s = 0$  clearly corresponds to an unpenalised prior and  $s = 1$  reduces to the original Equation (4). Also, setting  $s = -1$  encodes a PPP like Equation (6). This generalisation then raises the question of how should  $s$  be chosen? We suggest  $s = 1$  is a very natural choice, since this means the penalty is being applied on the original scale of the data. We have, however, looked at the sensitivity to the choice of  $s$ . For the example considered in Section 3, once  $s$  becomes large the posteriors for  $\mu$  become very flat.

## References

- Aitkin, M. (2001). Likelihood and Bayesian analysis of mixtures. *Statistical Modelling*, 1:287–304.
- Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, 39:1–38.
- Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. *Journal of the Royal Statistical Society, Series B*, 56:363–375.
- Fruhwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. Springer.
- Jasra, A., Holmes, C. C., and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modelling. *Statistical Science*, 20:50–67.
- Lee, K., Marin, J., Mengersen, K., and Robert, C. (2008). Bayesian inference on mixtures of distributions. *Handbook of Statistics*, 25(5):24.
- Marin, J. M., Mengersen, K. L., and Robert, C. P. (2005). *Bayesian modelling and inference on mixtures of distributions*. Elsevier.
- McLachlan, G. and Peel, D. (2000). *Finite Mixture Models*. Wiley.
- Mengersen, K. and Robert, C. (1996). *Bayesian Statistics*, chapter Testing for mixtures: a Bayesian entropic approach, pages 255–276. Oxford University Press, London.
- Nobile, A. and Fearnside, A. T. (2007). Bayesian finite mixtures with an unknown number of components: The allocation sampler. *Statistics and Computing*, 17(2):147–162.
- Pearson, K. (1894). Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society London, Series A*, 185:71–110.

- Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society, Series B*, 59:758–764. With discussion.
- Roeder, K. and Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, 92(439):894–902.
- Sperrin, M., Jaki, T., and Wit, E. (2010). Probabilistic relabelling strategies for the label switching problem in Bayesian mixture models. *Statistics and Computing*, 20(3):357–366.
- Stephens, M. (2000). Dealing with label-switching in mixture models. *Journal of the Royal Statistical Society, Series B*, 62:795–809.
- Wasserman, L. (2000). Asymptotic inference for mixture models using data-dependent priors. *Journal of the Royal Statistical Society, Series B*, 62:159–180.